Designing Statistical Estimators That Balance Sample Size, Risk, and Computational Cost

Tongliang Deng

SIST

2018-6-20

Motivation and general method

- Motivation
  When we have a **large amount of data**, we can exploit excess samples to decrease statistical risk, to decrease computational cost, or to trade off between the two.

- Method
  Smooth statistical optimization problems more and more aggressively as the amount of available data increases.

Introduction

- The Data Model

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^{\natural} + \boldsymbol{v}, \boldsymbol{A} \in \mathbb{R}^{mxd}$$

  Problem: $\hat{x} := \arg\min_x f(x), s.t. ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}|| \leq \sqrt{m \cdot R_{max}} =: \epsilon$

- Descent cone

  The descent cone of a proper convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ at point $\boldsymbol{x} \in \mathbb{R}^d$ is the convex cone

$$\mathcal{D}(f; x) := \bigcup_{\tau > 0} \{\boldsymbol{y} \in \mathbb{R}^d : f(\boldsymbol{x} + \tau \boldsymbol{y}) \leq f(\boldsymbol{x})\}$$

- Statistical dimension

  Let $\mathcal{C} \in \mathbb{R}^d$ be a closed convex cone. Its statistical dimension $\delta(\mathcal{C})$ is

$$\delta(\mathcal{C}) := \mathbb{E}_g[||\Pi_{\mathcal{C}}(g)||^2]$$

Introduction

- Phase Transition

  Whenever $m < \delta$,

  $$\lim_{\sigma \to 0} \frac{\mathbb{E}_v[R(x^\star)|\boldsymbol{A}]}{\sigma^2} = 0$$

  with probability p_1.

  Whenever $m > \delta$,

  $$|\lim_{\sigma \to 0} \frac{\mathbb{E}_v[R(x^\star)|\boldsymbol{A}]}{\sigma^2} - (1 - \frac{\delta}{m})| \leq t m^{-1} \sqrt{d}$$

  with probability p_2.

Geometric opportunity

Enlarging convex constraint sets can make corresponding statistical optimization problems easier to solve. These geometric deformations, however, create a loss of statistical accuracy.

- By enlarging the sublevel sets of the regularizer f , we increase the statistical dimension of the descent cone of f at $\boldsymbol{x}^\natural$.

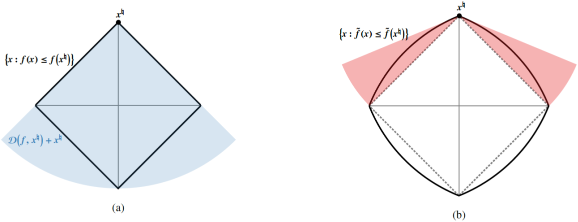


Fig. 1. **A geometric opportunity.** Panel (a) illustrates the sublevel set and descent cone of a regularizer f at the point $\boldsymbol{x}^\natural$. Panel (b) shows a relaxed regularizer $\tilde{f}$ with larger sublevel sets. The shaded area indicates the difference between the descent cones of $\tilde{f}$ and f at $\boldsymbol{x}^\natural$. Fact III.3 shows how this difference in the size of the descent cones translates into a difference in statistical accuracy. We may compensate for this loss of statistical accuracy by choosing a relaxation $\tilde{f}$ that allows us to solve the optimization problem faster.

Computational opportunity

- Primal problem
  $\hat{x}_\mu := \arg\min_x f_\mu(x)$, subject to$||\boldsymbol{Ax} - \boldsymbol{b}|| \leq \epsilon$
- Dual problem
  maximize $g_\mu(z) := \inf_x\{f_\mu(\boldsymbol{x}) - < z, \boldsymbol{Ax} - \boldsymbol{b} > -\epsilon||z||\}$,
  where z is the dual variable.
- Rewrite the dual problem

$$
\begin{aligned}
g_\mu(z) &= \inf_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) - \langle \boldsymbol{A}^T z,\ \boldsymbol{x} \rangle \right\} + \langle z,\ \boldsymbol{b} \rangle - \epsilon \, \|z\| \\
&= \underbrace{-f_\mu^*(\boldsymbol{A}^T z) + \langle z,\ \boldsymbol{b} \rangle}_{\tilde{g}_\mu(z)} \underbrace{- \epsilon \, \|z\|}_{h(z)},
\end{aligned}
$$

where $f_\mu^\star$ is the convex conjugate of f_μ. $\tilde{g}_\mu$ has a Lipschitz gradient $L_\mu = \mu^{-1}||A||^2$.

Auslender-TeBoulle Algorithm

Algorithm 1. **Auslender–Teboulle**

Input: measurement matrix $\boldsymbol{A}$, observed vector $\boldsymbol{b}$, parameter ϵ

1: $z_0 \leftarrow \mathbf{0}$, $\bar{z}_0 \leftarrow z_0$, $\theta_0 \leftarrow 1$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad \boldsymbol{y}_k \leftarrow (1 - \theta_k) z_k + \theta_k \bar{z}_k$
4: $\quad \boldsymbol{x}_k \leftarrow \arg\min_{\boldsymbol{x}} \{ f(\boldsymbol{x}) + \langle \boldsymbol{y}_k, \ \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \rangle \}$
5: $\quad \bar{z}_{k+1} \leftarrow \text{Shrink}\left(\bar{z}_k - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)/(L_\mu \cdot \theta), \epsilon/(L_\mu \cdot \theta) \right)$
6: $\quad z_{k+1} \leftarrow (1 - \theta_k) z_k + \theta_k \bar{z}_{k+1}$
7: $\quad \theta_{k+1} \leftarrow 2/(1 + (1 + 4/\theta_k^2)^{1/2})$
8: **end for**

There is an bug at line 5, it should be

$$\bar{z}_{k+1} \leftarrow Shrink(\bar{z}_k + (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)/(L_\mu \cdot \theta), \epsilon/(L_\mu \cdot \theta))$$

Time-date tradeoff

- A dual-smoothing method
  Given a regularizer f in the problem, introduce a family $\{f_\mu : \mu > 0\}$ of strong convex majorants:

$$f_\mu(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{\mu}{2}||\boldsymbol{x}||^2$$

  These majorants have larger sublevel sets, and their descent cones have larger statistical dimension.
- Choosing a smoothing parameter
  1. Constant smoothing, choose a constant value of μ.
  2. Constant risk, $\frac{\delta(\mathcal{D}(f_\mu;\boldsymbol{x}^\natural))}{m} = \frac{\bar{\delta}}{\bar{m}}$.
  3. A tunable balance, $\frac{\delta(\mathcal{D}(f_\mu;\boldsymbol{x}^\natural)}{m} = \frac{\bar{\delta}}{\bar{m}+(m-\bar{m})^\alpha}$.

Experimental setup

- Intel Core 5, memory 8GB, matlab2016Rb, Linux.
- Sparse vector regression.

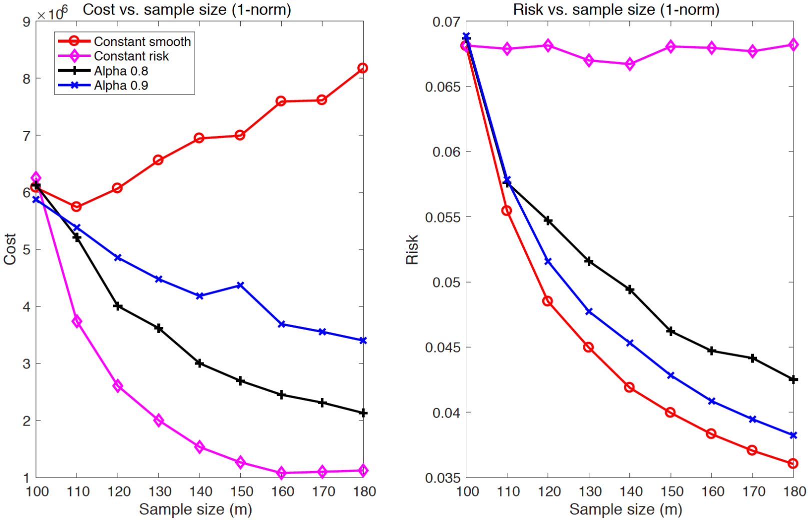
$$\delta(\mathcal{D}(f_\mu; x)) = d \cdot \psi(\rho)$$

where $\psi : [0, 1] \to \mathbb{R}$ *is the function given by*

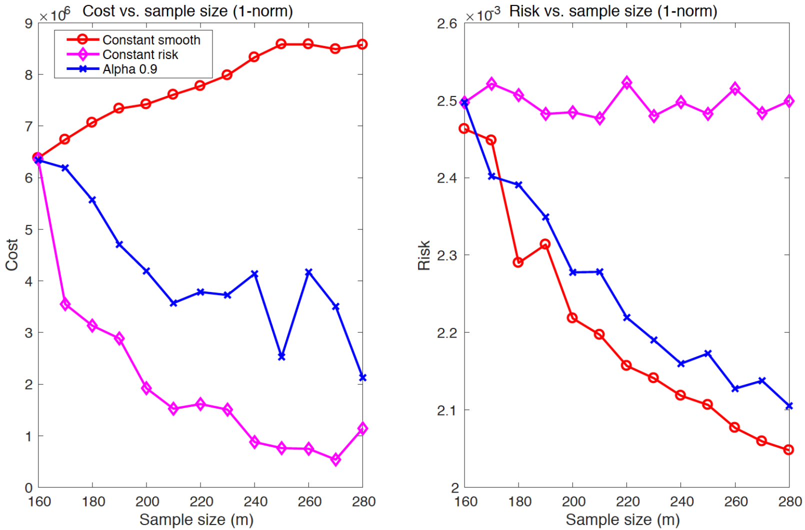
$$\psi(\rho) = \inf_{\tau \geq 0} \left\{ \rho \left[1 + \tau^2 (1 + \mu \|\boldsymbol{x}\|_{\ell_\infty})^2 \right] + (1 - \rho) \sqrt{\frac{2}{\pi}} \int_\tau^\infty (u - \tau)^2 \mathrm{e}^{-u^2/2} \, \mathrm{d}u \right\}.$$

- Low-rank matrix regression.
  Refer to the paper

Result of sparse vector regression



Result of Low-rank matrix regression



Thank you!